01｜声音是如何保存成数字信号的？

2021-11-22 冯建元

《搞定音频技术》

课程介绍 >

讲述：冯建元
时长 15:32　大小 14.24M

你好，我是建元。

作为专栏的第一节课，今天我们来聊聊音频信号中的一些基础知识，带你近距离地了解一下音频这个既熟悉又陌生的领域。这节课我们重点学习一下音频信号的关键指标和常见的音频封装格式，它们算是基础中的基础，希望你能够快速掌握。

音频信号的关键指标

声音我们每天都会听见，似乎早已习以为常。那么我们是怎么把声音信号转换成数字信号记录下来存储和传输的呢？

声音是听觉对声波产生的感知，而声波的本质是介质的振动，比如空气的振动。那么我们只需要把这个振动信号记录下来，并用一串数字来表达振动信号振动的快慢和振动的幅度，就可以实现声音的记录。

如图 1 所示，以前的留声机就是通过唱片上凹槽的深浅、长短来表征声音的振幅和持续时间。

图1 留声机与唱片

我们现在一般用麦克风来实现声音的采集。**那如何通过麦克风来采集声音呢？**

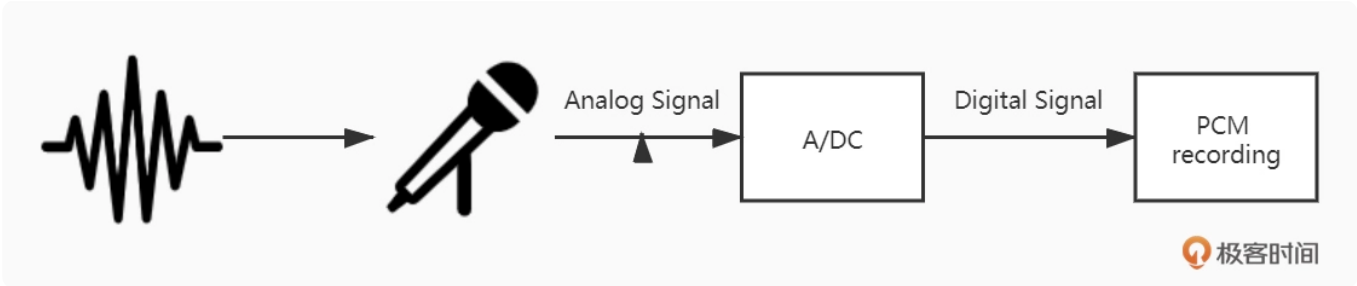

图2 音频数字信号的生成

使用麦克风的音频数字信号采集过程如图 2 所示：

首先，声波通过空气传播到麦克风的振膜。

然后，振膜随空气抖动的振幅大小产生相应的电学信号。**我们把这种带有声学表征的电学信号叫做模拟信号（Analog Signal）。**

最后，**通过 A/DC（模数转换器）将模拟信号转换成数字信号（Digital Signal）**。即通过 PCM（Pulse Code Modulation）脉冲编码调制对连续变化的模拟信号进行抽样、量化和编码转换成离散的数字信号。

这样我们就实现了音频信号的采集，我们常说的 PCM 文件就是未经封装的音频原始文件或者叫做音频“裸数据”。那么具体音频的数字信号是怎么构成的呢？这就涉及到下面的 3 个基本概念：采样位深、采样率和通道数。

现在我们先来熟悉一下这 3 个概念。

采样位深

采样位深也就是每个采样点用多少 bit 来表示。比如位深是 16 就代表每个采样点需要 16bit 来进行存储。从物理意义上来说，**位深代表的是振动幅度的表达精确程度或者说粒度**。假设数字信号是一个 1 到 -1 的区间，如果位深为 16bit，那么第 1 个 bit 表示正负号，并且剩下的 15 个 bit 可以表征 0～32767 个数，那么振幅就可以精确到 1/32768 的粒度了。

我们一般在网络电话中用的就是 16bit 的位深，这样不太会影响听感，并且存储和传输的耗费也不是很大。而在做音乐或者更高保真度要求的场景中则可以使用 32bit 甚至 64bit 的位深来减少失真。

8bit 时失真就比较严重了。早期受到音频技术条件限制，很多音频都是 8bit 的，声音会显得比较模糊，如今也只有一些电话和对讲机等设备还有使用。但有趣的是，有的音乐就追求这种模糊感，所以“8bit”有的时候也代表一种听感朦胧的音乐艺术类型。

采样率

采样率就是 1 秒内采集到的采样点的个数，一般用赫兹 Hz 来表示。比如 1 秒有 48000 个采样点那么采样率就是 48000Hz（48kHz）。

根据奈奎斯特采样定理在进行模拟 / 数字信号的转换过程中，当采样频率 fs 大于信号中最高频率 fmax 的 2 倍时（fs > 2fmax），采样之后的数字信号才可以完整地保留原始信号中的信息。也就是说采样率和保留的声音频率基本上是 2 倍的关系。

我们可以看看图 3 的频谱图来对比一下 16kHz 采样率和 48kHz 采样率的音频。

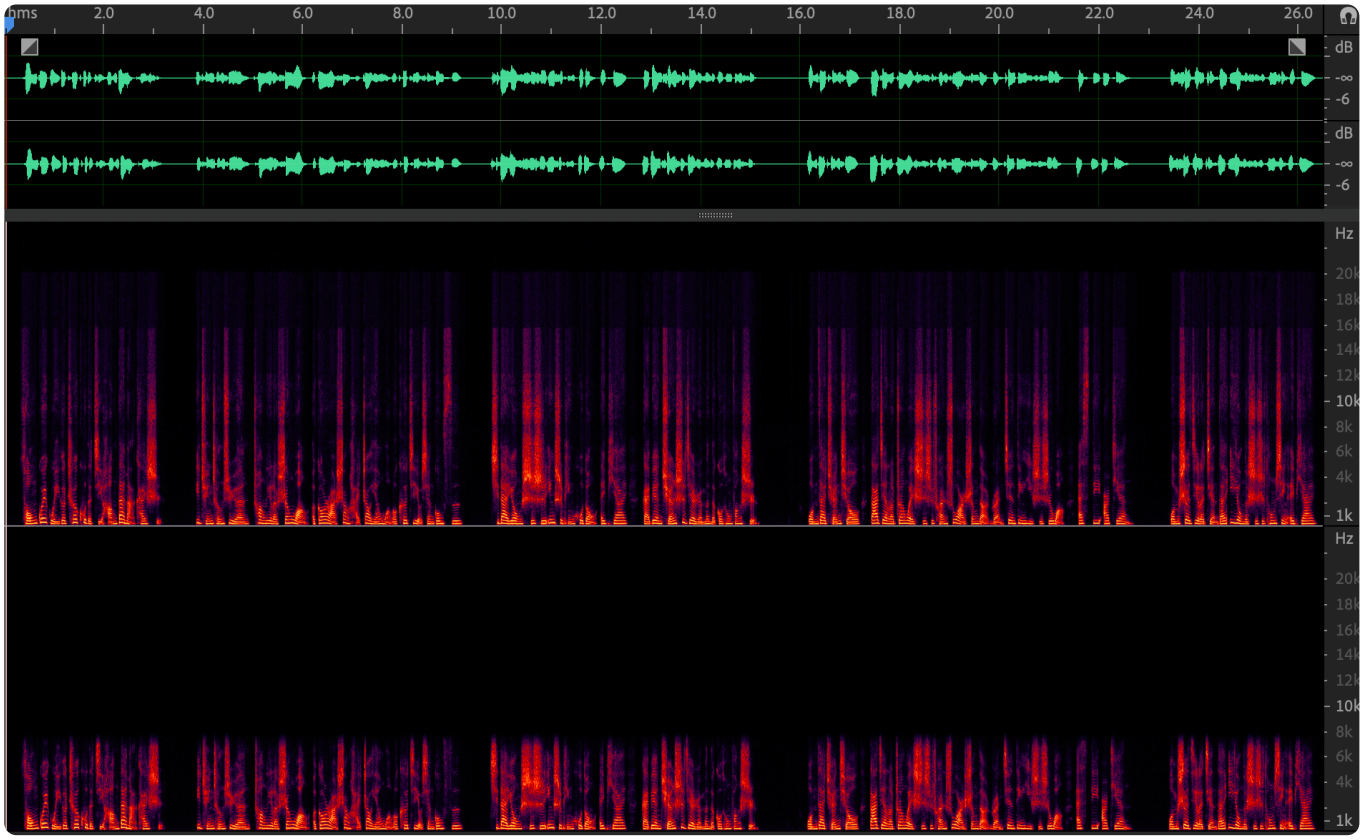

图3 不同采样率（ 48kHz(上)和16kHz(下) ）的频谱能量分布

由图 3 我们可以看到，16kHz 采样率的音频在 8kHz 以上的频谱基本是没有能量的（黑色），也就是说这部分高频的信息由于采样率不够已经丢失了。从听感上来说人耳可以听到的频率范围大概是 50～20kHz 之间。如果采样率不够，那么和实际听感比起来声音就会显得“低沉”或者说“闷”。

那么采样率是不是越高越好呢？其实选用什么样的采样率是根据具体用途来决定的。

如果只是为了听见人声、听懂对方在说什么，那么为了节省传输码率我们可以把采样率降到 8kHz（比如打电话）。而在网络音视频会议场景需要平衡音质和传输带宽消耗，我们一般可以使用 16kHz 或者 32kHz 的采样率。如果是开线上音乐会或者音乐直播，我们通常会用较高的采样率来保证音质，比如 44.1kHz 或者 48kHz。更极端一点，在音乐制作录音的时候，我们会采用 96kHz 甚至更高的采样率来方便后续的调音和制作。

通道数

你可能在平时买音响的时候听过 2.1 声道或者 5.1 声道等名词，这些数字代表了有多少个播放单元。比如，2.1 声道中的 2 指的是左右两个音箱，1 指的是中间一个低音音箱（如图

4 所示）。每个音箱都会播放一个单独的音频，这时候就需要同时有 3 路音频信号同时播放，或者叫通道数为 3。

图4 2.0声道音箱(左)、立体声耳机(中)、2.1声道音箱(右)

我们在实时在线互动的时候，由于编 / 解码器能力的限制（比如使用了单通道编 / 解码器），或者采集设备能力的限制（只能采集单通道的信号），音频信号通常为单声道的。我们听歌的时候，戴上耳机如果听到左右耳朵是不一样的，能够感觉到声音是从不同方向传过来的，那么就说明这个音频是双声道。我们通常也把这种双声道音频叫做立体声（Stereo）（如图 4 所示）。

除了播放需要多声道以外，采集也可能采集到多通道的数据。比如麦克风阵列采集到的原始信号，有多少个麦克风就会有多少个通道的音频信号。因此，**这里通道数的物理含义其实就是同一时间采集或播放的音频信号的总数。**

好了，这里我们已经基本掌握了音频数字信号的基本构成。现在我们来通过一个简单的计算来理解一下。假设我们有一个立体声的 PCM 音乐文件。它记录了 1 分 40 秒的采样率为 48kHz 的音频。如果这个文件的采样位深是 16bit，那么这个立体声文件应该占用多大的存储空间呢？如果不经过压缩实时传输播放，又至少需要多少的带宽呢？

有了前面的知识我们就可以知道，一个 PCM 音频文件的存储大小就是采样位深、采样率、通道数和持续时间的累乘。即：

$$存储空间 = 采样位深 \times 采样率 \times 通道数 \times 时长 = 16 \times 48000 \times 2 \times 100$$

而它实时传输所需的带宽就是它每秒所需的比特带宽。这可以用采样位深、采样率和通道数来得到。即：

$$比特带宽 = 采样位深 \times 采样率 \times 通道数 = 16 \times 48000 \times 2 = 1.53 * 10^6 bp$$

音频的封装

我们可以看到上面例子中 100 秒的 PCM 音频文件就需要这么大的存储空间，那我们平时经常看到的音频文件格式，比如 MP3、FLAC 和 WAV 等，它们有什么区别？它们所需的存储空间一样么？

有损和无损音频编码封装格式

其实就算不是音频行业的人，也应该或多或少地听过有损和无损的音频编码封装格式。顾名思义，有损的音频封装格式主要是通过压缩算法把文件大小尽量减少，但是在解压缩的时候却无法完美还原音频原来的数据（即有损）。比如 MP3、AAC、AMR 和 WMA 等编码封装格式。

虽然叫做有损音频格式，但其实发展到现在，有损音频格式比如 MP3 一般可以达到 1:10 的压缩比，即存储体积为未压缩音频的十分之一。但在听感上和无损格式比起来，如果不是专业人士很难听出区别。

而无损音频封装则采用可完美还原的压缩算法，比如 FLAC 和 APE 等编码封装格式。FLAC 与 APE 的压缩比基本相同，其中 FLAC 的压缩比为 58.70%，而 APE 的压缩能力则要更高一些，压缩比为 55.50%。它们都能压缩到接近源文件一半大小。无损封装甚至还可以不压缩编码，直接加个文件头作为封装，比如 WAVE 格式的封装。

其实简单地说，不是为了追求极致的听感，比如听音乐、音频分析，我们可以用有损压缩来节省存储体积。但是如果是拿音频信号做数据分析或者追求极致的听感还原，则必须使用无损压缩的封装格式来避免编码带来的损伤。

常见的有损和无损封装格式如下表 1 所示。其实很多编码封装也是支持调节压缩比来平衡音质和存储空间的，比如 WMA 可以同时支持无损和有损编码，而且 MP3 也可以调整不同的编码码率来调节音质。

无损编码封装	有损编码封装
APE、WAV、FLAC、AIFF、ALAC、WMA	MP3/MP3 Pro、AAC、AMR、WMA、RA、OGG

极客时间

表1 常见的无损和有损音频编码封装

因为无需封装、方便存储，所以音频工程师经常把音频处理的中间变量（未封装的音频数据）存储为 PCM 文件。但由于 PCM 文件也就是裸数据文件是无法直接读取播放的。比如把 PCM 文件拖到音频处理软件 Audacity 中，会提示你先输入这个 PCM 文件的采样率、通道数、比特位深等信息才能进行解析播放。这样，一个不知道这些音频信息的人就无法正确地解析播放这些 PCM 文件了。

接下来我们主要介绍一个常用的 WAVE 格式的封装，这样我们在日常处理音频裸数据的时候就可以将其封装。别人在查看、播放时，就可以直接用播放软件打开了。

WAVE 文件的封装

WAVE 文件作为多媒体中使用的声波文件格式之一，文件后缀名为 wav。它是以 RIFF 格式为标准的，RIFF 是英文 Resource Interchange File Format 的缩写。因此，每个 WAVE 文件的头四个字节便是“RIFF”。

WAVE 文件的封装格式十分简单。WAVE 文件由 WAVE 文件头部分和 WAVE 文件数据体部分组成，其中 0～43 字节存放采样率、通道数、数据部分的标识符等头信息，44 字节以后的就是数据部分。简单地理解就是 PCM 文件加一个文件头描述文件的基本信息。具体文件头每个字节的含义可以参考表 2。

偏移地址	字节数	数据类型	内容
0x00	4	char	“RIFF”标志
0x04	4	long int	文件长度
0x08	4	char	“WAVE”标志
0x0C	4	char	“fmt ”标志
0x10	4		过度字节（不定）
0x14	2	int	格式类别（0x10是PCM形式的声音格式）
0x16	2	int	通道数（1为单声道，2为双声道）
0x18	2	int	采样率（每秒的样本数，表示每个通道的播放速度）
0x1C	4	long int	波形音频数据传输速率（通道数x每秒数据位数x样本的数据位数/8，播放器根据此值估算缓冲区大小）
0x20	2	int	采样帧大小。数值为：通道数x位数/8。播放软件需要一次处理多个该值大小的字节数据，用该数值调整缓冲区。
0x22	2	int	每样本的数据位数，表示每个声道中各样本的数据位数（8bit或者16bit），如果有多个声道，对每个声道而言，样本大小都一样。
0x24	4	int	ASCII字符“data”，标示头结束，开始数据区域。
0x28	4	long int	数据大小

极客时间

表2 WAVE文件头每个字节的具体含义

从表格 2 中我们可以看到，WAVE 格式支持单、双声道的音频文件的封装，以及采样位深为 8bit 和 16bit 这两种格式。那么具体单通道和双通道数据是如何排列的呢？我们可以看一下图 5。

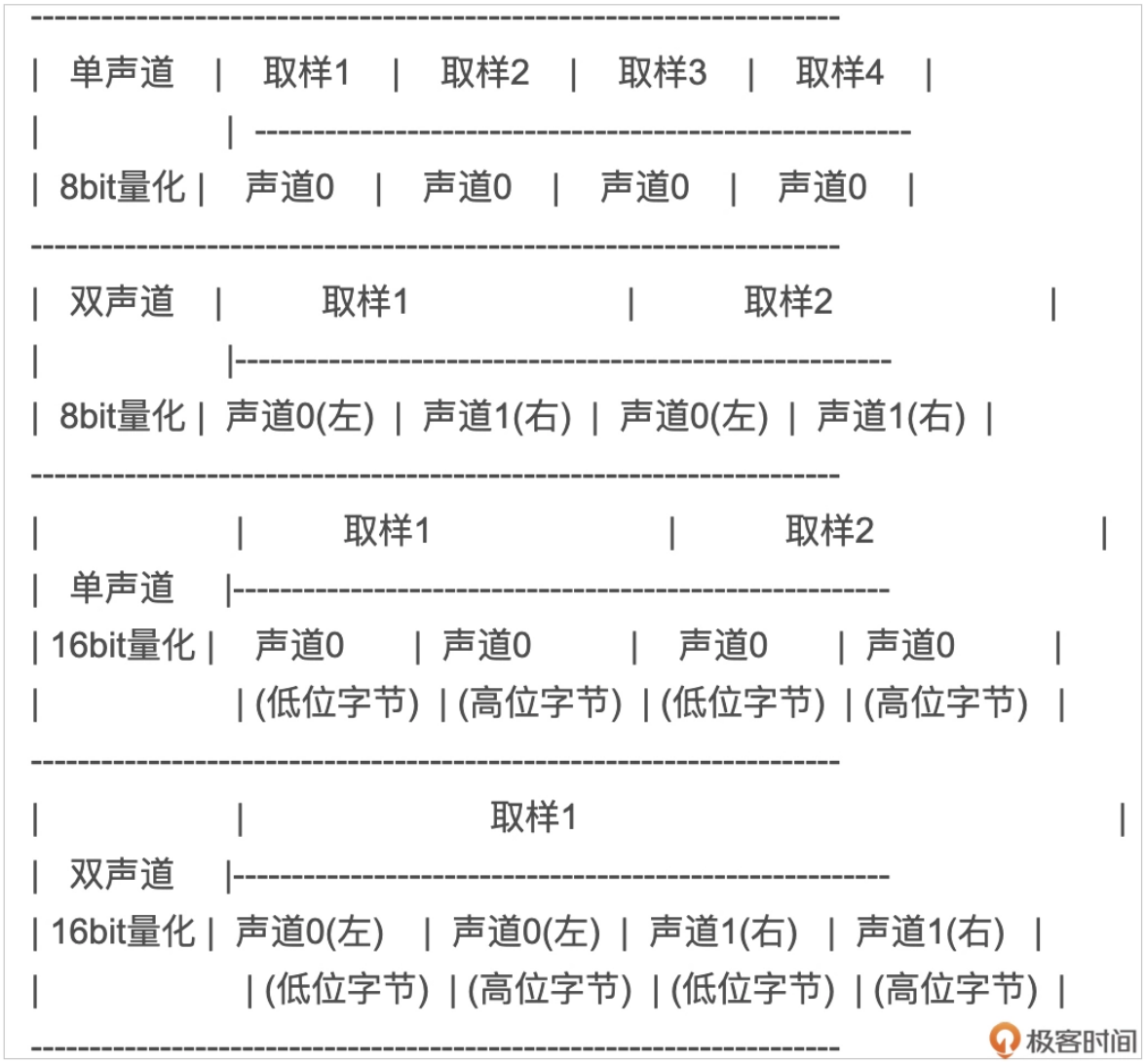
单声道	取样1	取样2	取样3	取样4
8bit量化	声道0	声道0	声道0	声道0
双声道	取样1	取样2
8bit量化	声道0(左)	声道1(右)	声道0(左)	声道1(右)
单声道	取样1	取样2
16bit量化	声道0	声道0	声道0	声道0
	(低位字节)	(高位字节)	(低位字节)	(高位字节)
双声道	取样1
16bit量化	声道0(左)	声道0(左)	声道1(右)	声道1(右)
	(低位字节)	(高位字节)	(低位字节)	(高位字节)
极客时间

图5 WAVE格式中数据的比特分布

从图 5 我们可以看到双声道的时候，是左右声道的数据间隔排布的，这主要是为了方便可以按照时间连续成块地读取数据信息，而不是读完一个声道再读下一个。

小结

好了，我们今天的课程到这里就要结束了。我们先来简单回顾一下今天讲到的知识点。

首先，采样位深、采样率、通道数是描述音频信号的关键指标。采样位深越深、采样率越高，则音质越好，但同时消耗的存储和传输的资源也就越多。

其次，在编码封装时如果需要尽量缩小存储音频的体积，那么可以选用 MP3 等有损的编码封装格式，而需要无损、高保真的音频时，则可以采用例如 FLAC 等无损格式来编码封装。其中对于音频开发者而言，WAVE 格式由于可以迅速快捷地封装一个音频裸数据文件而被广泛使用。

最后，在实际使用时我们就可以根据自己的使用场景来选择音频信号的关键指标和封装形式了。

记住，音频信号主要是用来给人听的，在音频处理时往往需要“眼看为虚，耳听为实”。你可以多听听不同的采样率，不同的采样位深会有什么样的听感。听着听着就会形成记忆，以后解决问题时，你甚至只需要听一耳朵就能发现问题。

思考题

学习了上面的知识点，现在我有一道思考题留给你。

为什么有的音频文件，文件显示的是 48kHz 的采样率和 16bit 的位深，但听起来还是很闷或者听感很差呢？

欢迎你在留言区和我分享你的思考和疑感，你也可以把今天所学分享给身边的朋友，邀请他加入探讨，共同进步。下节课再见。

分享给需要的人，Ta订阅后你可得 20 元现金奖励

生成海报并分享

赞 3 提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 开篇词 | 实时互动强势发展，如何快速入门音频技术？

下一篇 02 | 如何量化分析语音信号？

精选留言 (13)

写留言

徐刚
2021-11-26

思考题：

从音频文件采集，存储，播放来看，任何一个环节都会引发这个问题；
采集：原始音频文件采集的设备质量不高，导致高频段有损伤
存储：音频文件采用有损压缩，可能会造成高频段损伤
播放：播放设备质量差，不能很好表现高频段信号

展开

作者回复: 这个回答的很全面了。

3

springXu
2021-11-22

关于思考题
原始的采集设备只是普通话筒，再高的采样率和位深，也是没用的。又或者输出的喇叭只是能还原低频部分音效。

问个老技术...

展开

作者回复: 这位同学回答的不错。MIDI全称是Musical Instrument Digital Interface顾名思义是一种音乐乐器的记录格式。以钢琴为例，MIDI文件记录了时间序列中每个钢琴音的起止位置、按键力度等信息。很显然光有MIDI文件你是听不到声音的还需要有一个记录了乐器各种音的音调的音源库配套使用，比如文件记录在第3秒要弹一个C4的音，然后就去音源库把C4的音源按时间拼接上去，最后依次把所有的乐器音都拼接起来就成了可播放的音频文件。起初各大厂商MIDI文件格式都不太统一。直到后来MIDI1.0标准的发布各家厂商遍都遵循了相同的标准。现在已经发展到MIDI2.0了。

共 2 条评论 >

3

ripple
2021-11-23

重采样吧

展开

作者回复: 没错，这是其中的一种很常见情况，音频直接上采样用的是插值的方法，是无法恢复高频信息的。如果原始音频文件的质量很差或者采集设备的音质很差那么即使这些采样率的指标看起来很好音质也不会好听。

2

不负
2021-12-04

思考题：从“1.音频采集到PCM-2.音频封装存储-3.音频播放”的整个流程来看，某个环节设备出现问题或不匹配，抑或是编码封装存储时音频受损，也可能不是对原始音频进行采样

「捉个虫」关于 WAVE 文件头信息说明那段，表格里 0x18 对应的应该是 4个字节吧（0x18～0x2C）

展开

作者回复: 是的，多谢指出。

梁智行
2021-12-03

位深相当于用多少维的embedding表达一个字，采样率相当于一句话的 sequence length，通道数相当于图像的通道那rgb维度，时间就是就是batch大小，原来音乐是一个四维数据的结构。

pencilCool
2021-11-28

思考题：
原始音频不清晰，采样率再高也没用 。

大土豆
2021-11-26

老师，关键问题来了，16bit第一位表示正负号，不也可以区分振幅吗？可以表示65535个振幅吧？

作者回复: 这里说的振幅也就是振动幅度，是一个大于0的值。如果理解为采样值就可以有正负了。

Geek_7de4c5
2021-11-25

这种问题大部分由于设备差引起的，请问老师有办法修复吗？将录制差的音质变好呢？

作者回复: 这就要看是什么样的情况，比如人声被降采样没有高频，这时候有些方法比如频谱扩展就可以使用。但有的时候比如音乐的录制想要做一些恢复就往往没那么容易。

yu
2021-11-24

老师，是不是单通道的音频，采样源就是一个麦克风(打了比方，可能不是麦克风)，立体声就是放左右两个麦克风，2.1就是两个麦克风加一个专门录低音的设备?

作者回复: 立体声采集是需要两个麦克风，但这里说的2.1其实指的是"播放设备"需要3个扬声器(左、右、中)。采集设备一般不用2.1、5.1这种来表示。

liukeying
2021-11-24

看了文章中有关WAVE文件封装格式部分，其中"采样帧大小的计算是：通道数 * 位数 / 8"，没明白到底是什么含义，套入数据，也没计算出有用的数据。还请老师给讲一下。

作者回复: 这里主要是看同一时间需要读取多少个字节，一个字节是8bit，所以要除以8.同时要播放所有声道的采样点，所以举个例子一个双声道16bit位深，需要一次读入2*16/8=4byte。

共 2 条评论 >

IT小僧
2021-11-24

老师问个问题，文章中提到: 从听感上来说人耳可以听到的频率范围大概是 50 ~ 20kHz 之间。 那岂不是48khz听不到啥有用的信息了。这个50单位是Hz还是KHz

作者回复: 50是Hz哦，如果音频频率太高人耳确实就听不到什么信息了，也就是所谓的超声波。48kHz的采样率包含了24kHz一下的所有声音频段的信息不只是一个频率的信息。不同人的耳朵还有些区别，比如20kHz-24kHz有的人能听见，有的则不行。

共 2 条评论 >

lckfa李钊
2021-11-24

请教老师一个问题，声音在渲染前，是否要求字节对齐呢？

作者回复: 音频中每个采样点都需要按照顺序排列播放，不然人耳很容易听到杂音异常。

T------T

2021-11-22

思考题

1.这个文件是由一个8000kHz 转换过来的

2.音频采集的时候声源本身就是分布在低频段的声音

展开

作者回复: 回答的很对！再举个例子，也有的时候音源很好，但播放设备不好也有可能，比如手机的扬声器被堵住了。